Bioinformatics, 38(22), 2022, 5081–5091
https://doi.org/10.1093/bioinformatics/btac639
Advance Access Publication Date: 20 September 2022
Original Paper

Data and text mining

Reproducible acquisition, management and meta-analysis of nucleotide sequence (meta)data using q2-fondue

Michal Ziemski [†], **Anja Adamov** [†], **Lina Kim, Lena Flörl and Nicholas A. Bokulich***

Laboratory of Food Systems Biotechnology, Institute of Food, Nutrition, and Health, ETH Zürich, Zürich 8092, Switzerland

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Jonathan Wren

Received on May 4, 2022; revised on September 8, 2022; editorial decision on September 9, 2022; accepted on September 19, 2022

Abstract

Motivation: The volume of public nucleotide sequence data has blossomed over the past two decades and is ripe for re- and meta-analyses to enable novel discoveries. However, reproducible re-use and management of sequence datasets and associated metadata remain critical challenges. We created the open source Python package *q2-fondue* to enable user-friendly acquisition, re-use and management of public sequence (meta)data while adhering to open data principles.

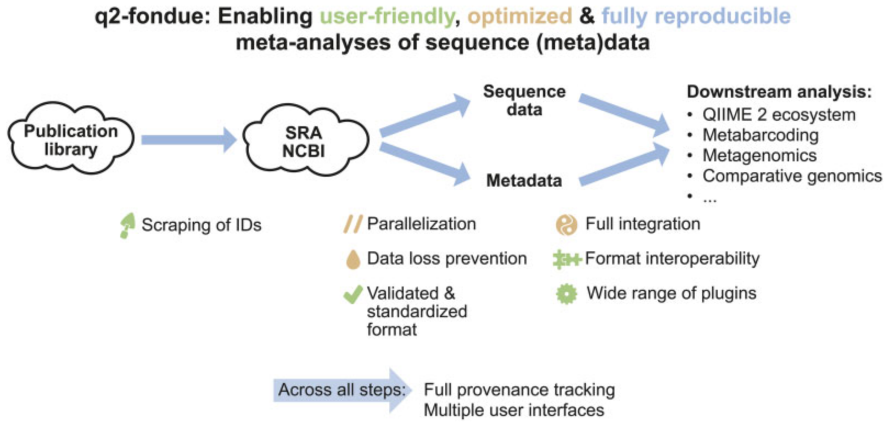
Results: *q2-fondue* allows fully provenance-tracked programmatic access to and management of data from the NCBI Sequence Read Archive (SRA). Unlike other packages allowing download of sequence data from the SRA, *q2-fondue* enables full data provenance tracking from data download to final visualization, integrates with the QIIME 2 ecosystem, prevents data loss upon space exhaustion and allows download of (meta)data given a publication library. To highlight its manifold capabilities, we present executable demonstrations using publicly available amplicon, whole genome and metagenome datasets.

Availability and implementation: *q2-fondue* is available as an open-source BSD-3-licensed Python package at https://github.com/bokulich-lab/q2-fondue. Usage tutorials are available in the same repository. All Jupyter notebooks used in this article are available under https://github.com/bokulich-lab/q2-fondue-examples.

Contact: nicholas.bokulich@hest.ethz.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Graphical Abstract



1 Introduction

The increasing volume of publicly available nucleotide sequence data is driving a revolution in the life sciences, by enabling comparative studies to discover generalizable trends that are often inaccessible or underpowered in an individual study. Individual studies addressing similar biological questions can differ in many technical aspects, including (but not limited to) specific experimental design, employed sequencing technologies, definitions of the examined target variables and selection of potential covariates influencing the target. These inter-study variations can make individual study results biased (Serghiou *et al.*, 2016) and even contradictory to one another (Ioannidis and Trikalinos, 2005). Meta-analysis allows the synthesis of findings from individual studies to reach a more complete understanding: identifying consistent and reproducible signatures across studies and resolving causes of variation among study results (Gurevitch *et al.*, 2018).

Meta-analyses of nucleotide sequencing-based studies have intensified within the past decade (see Fig. 1), given the high potential of these data for reuse in comparative analyses. Meta-analyses of genome-wide association data have expanded our knowledge of human polygenic disorders and quantitative traits (Panagiotou *et al.*, 2013). Comparative genomics has given insights into vertebrate genome evolution (Meadows and Lindblad-Toh, 2017) and the processes of genome function, speciation, selection and adaptation (Alföldi and Lindblad-Toh, 2013). Comparative analyses of global microbiome datasets have driven deepening insight into spatiotemporal and biogeographic variation in Earth's microbial diversity (Tara Oceans Consortium Coordinators *et al.*, 2015; Thompson *et al.*, 2017). Re-analysis of the published genome and metagenome data has enabled the discovery of novel and candidate microbial clades, as in the Genome Taxonomy Database (Parks *et al.*, 2017), and highlighted the abundance (Lloyd *et al.*, 2018) and ecosystem-impact (Zamkovaya *et al.*, 2021) of uncultured microbes, also known as 'microbial dark matter'. Since meta-analyses can only be conducted if the original study data are publicly available, the recent increase in meta-analyses can be partly attributed to the ongoing open-science efforts of making sequencing data and accompanying metadata standardized and publicly available (Berman *et al.*, 2014; McNutt *et al.*, 2016; The Path to Open Data, 2019; Yilmaz *et al.*, 2011). The Sequence Read Archive (SRA), established as part of the International Nucleotide Sequence Database Collaboration (INSDC) by the National Center for Biotechnology Information (NCBI), enables free access to sequence data (Kodama *et al.*, 2012; Leinonen *et al.*, 2011b), including sequences stored on ENA (Leinonen *et al.*, 2011a) and DDBJ (Mashima *et al.*, 2017). Since its creation in 2009, the SRA has gathered data at the petabyte scale and continues to scale its infrastructure to ensure efficient data storage and retrieval (Katz *et al.*, 2021).

A selection of tools to programmatically access data from the SRA has recently emerged. NCBI offers the *sra-tools* command-line toolkit (Leinonen *et al.*, 2011b) for downloading and interacting with raw sequence data. The *entrezpy* Python library (Buchmann and Holmes, 2019) aids in automating the data download from NCBI's Entrez databases by providing abstract classes allowing custom implementations. *pysradb* (Choudhary, 2019) makes use of the curated metadata database available through the SRAdb project (Zhu *et al.*, 2013) to download data from the SRA, and *grabseqs* enables fetching of data from not only the SRA but also MG-RAST (Meyer *et al.*, 2008) and iMicrobe (Youens-Clark *et al.*, 2019). However, the wider application of these tools in large comparative analyses is hindered by several challenges, including technical complication and a steep learning curve for users with limited programming skills, and the difficulty of tracking data provenance necessary for reproducibility and traceability.

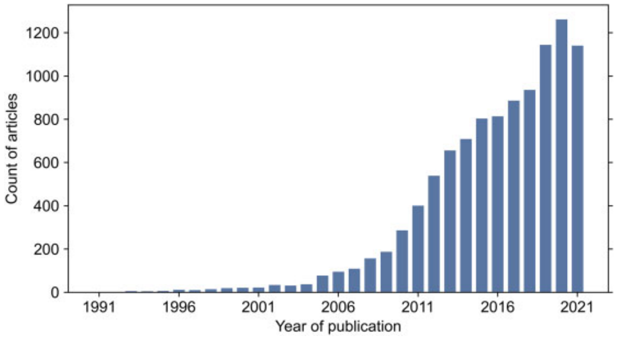


Fig. 1. Increasing trend of sequencing-based meta-analyses over the past 30 years. Displayed article counts were retrieved from PubMed on February 21, 2022 with the search query '(meta-analysis) AND ((omics) OR (genom*) OR (microbio*) OR (transcriptom*))' and a requirement for the article type to be a meta-analysis

In order to provide consistent and reliable findings, meta-analyses must follow Findable, Accessible, Interoperable and Reusable (FAIR) Guiding Principles (Wilkinson *et al.*, 2016). To this end, meta-analyses should be performed in a reproducible manner, making use of consistent workflows while keeping track of all the performed data retrieval and analysis steps. Despite increasingly facilitated access to sequencing data, reproducibility and provenance of primary and secondary studies remain challenging (Amann *et al.*, 2019; Baker, 2016; Huang and Gottardo, 2013; Kim *et al.*, 2018; Reichman *et al.*, 2011). Here, we introduce an open-source software package *q2-fondue* (**F**unctions for reproducibly **O**btaining and **N**ormalizing **D**ata re-**U**sed from **E**lsewhere) to expedite the initial acquisition of data from the SRA, while offering complete provenance tracking. *q2-fondue* simplifies retrieval of sequencing data and accompanying metadata in a validated and standardized format interoperable with the QIIME 2 ecosystem (Bolyen *et al.*, 2019). By allowing access through multiple QIIME 2 user interfaces, it can be employed by users of different computational capabilities.

Here, we describe the *q2-fondue* software package and subsequently demonstrate its use in comparative analyses of marker gene, genome and metagenome sequencing studies. We anticipate *q2-fondue* will lower existing barriers to comparative analyses of nucleotide sequence data, facilitating more transparent, open and reproducible conduct of meta-analyses.

2 Implementation

2.1 Software overview

q2-fondue is an open source Python 3 package released under the BSD 3-clause license. It can be installed in a conda environment on any UNIX-based system as described in the installation instructions provided on the package website (https://github.com/bokulich-lab/q2-fondue). *q2-fondue* has been implemented as a QIIME 2 plugin, allowing the use of QIIME 2's integrated data provenance tracking system, multiple user interfaces and streamlined interoperability with downstream sequence analysis plugins.

An overview of *q2-fondue* is shown in Figure 2. Two separate *q2-fondue* actions allow easy access to the SRA database: *get-sequences* and *get-metadata* fetch per-sample sequence data and corresponding metadata (e.g. sample and run information), respectively. The *get-all* pipeline wraps both of these actions to simultaneously retrieve sequences and metadata for a list of SRA accessions. These three actions, *get-sequences*, *get-metadata* and *get-all*, require a single input file containing accession IDs of one type to be fetched. Currently, *q2-fondue* supports BioProject, study, sample, experiment and run accession IDs. BioProject and study IDs are in a one-to-one relationship, where a study ID denotes the SRA record of the associated BioProject ID. All other IDs are in a one-to-many relationship in the listed order (e.g. one study ID, with its linked BioProject ID, can have many sample IDs associated with it—see Fig. 3 for an overview). Run IDs allow direct interaction with the SRA databases, while the other IDs are first translated into corresponding run IDs using a chain of E-Direct utilities (Kans, 2013; https://www.ncbi.nlm.nih.gov/books/NBK179288/). An *E-Search* query is executed to look up provided IDs in the BioProject (if BioProject ID was provided as input) or SRA (if study, sample or experiment ID was provided as input) database, followed by an

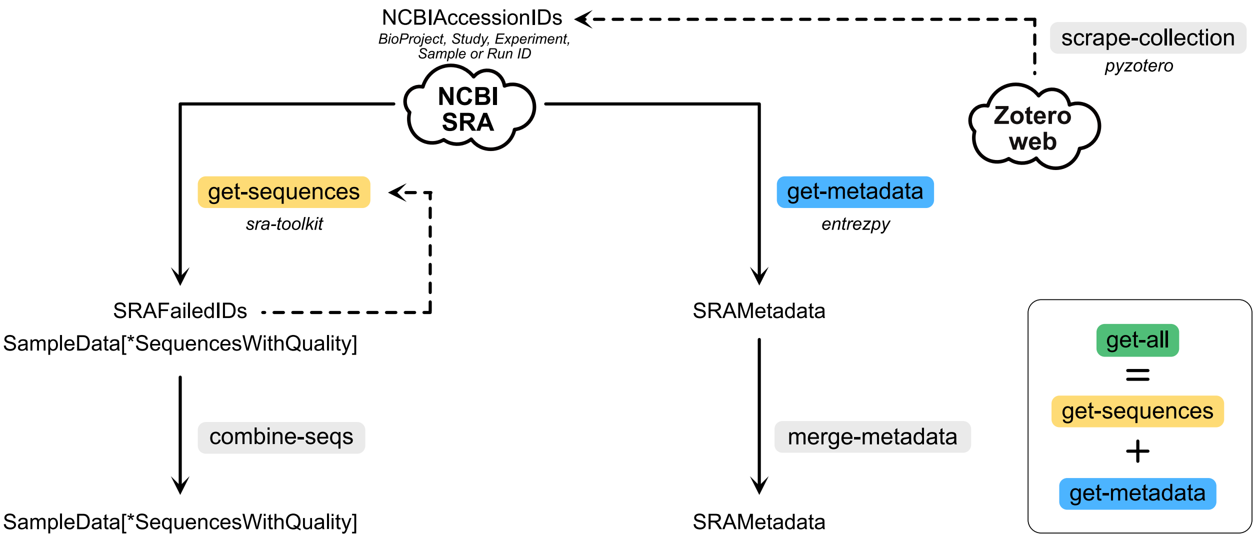


Fig. 2. Overview of *q2-fondue* methods. *get-sequences* method can be used to fetch raw sequencing data (single- and/or paired-end), while *get-metadata* can download corresponding metadata. Both methods can be run simultaneously by using the *get-all* pipeline, which produces all outputs (SampleData, SRAFailedIDs and SRAMetadata). *get-sequences*, *get-metadata* and *get-all* are run with a list of one type of accession IDs (BioProject, study, sample, experiment or run ID). Sequences obtained from multiple fetches can be combined using *combine-seqs* and multiple metadata artifacts can be merged with *merge-metadata*. All accession IDs can be retrieved from Zotero web library collections with the *scrape-collection* action

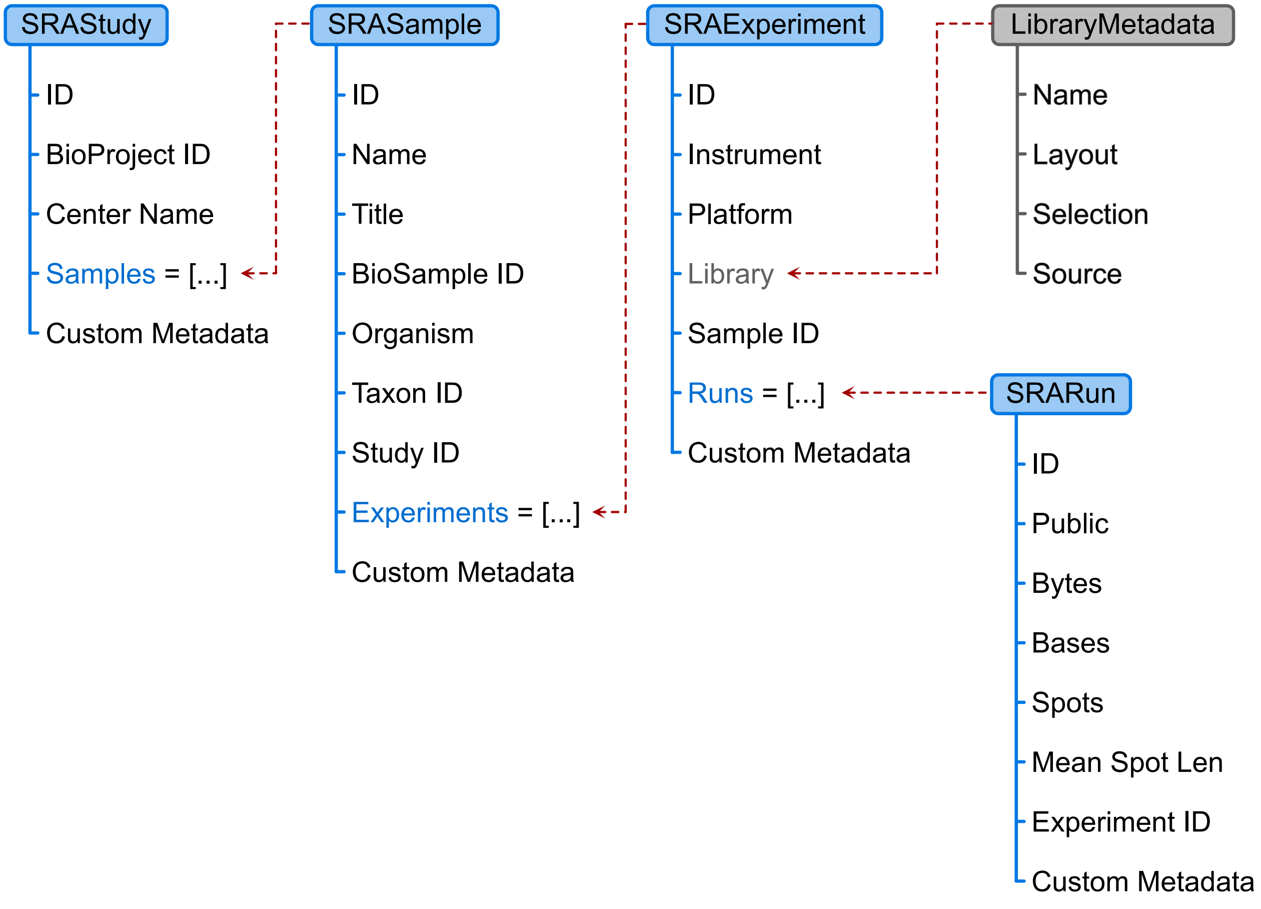


Fig. 3. Structure of the SRA metadata data classes used by the *get-metadata* action. Each of the classes represents a different level of metadata organization and can contain other nested metadata objects. As all the objects are linked together, metadata of the entire study and all its children can be retrieved directly from the SRAStudy object

E-Link query finalized by an *E-Fetch* query to retrieve the linked run IDs.

All data-fetching actions support configurable parallelization to maximally reduce the processing time. The *get-metadata* method employs a multi-threading approach built into the *Entrezpy* modules (see Section 2.2.), while *get-sequences* uses multiple processes and queues to coordinate the data download with its pre- and post-processing steps. Note that parallelization does not improve download speed (for which network bandwidth is the limiting factor), but markedly decreases pre- and post-processing runtime (e.g. data validation steps) (Supplementary Fig. S1).

2.2 Sequence retrieval

The *get-sequences* action makes use of the *sra-tools* command-line toolkit (Leinonen *et al.*, 2011b; https://github.com/ncbi/sra-tools) developed by NCBI. The *prefetch* tool is first used to reliably fetch SRA datafiles using the provided run IDs and the *fasterq-dump* utility is then executed to retrieve the corresponding sequences (single- or paired-end) in the FASTQ format. To follow QIIME 2's naming convention those sequences are then renamed using their accession IDs, compressed and finally validated by QIIME 2's built-in type validation system. The action keeps track of any errors that occurred while fetching the sequences and performs available storage checks on every iteration to ensure no data are lost when space is exhausted during download. There are three output files generated by the *get-sequences* action: two QIIME artifacts corresponding to single- and paired-end reads, respectively, and one table containing the list of IDs for which the download failed (if any) including the linked error messages.

2.2 Metadata retrieval

Retrieval of SRA metadata is possible through the *get-metadata* action. This action uses the *Entrezpy* package (Buchmann and Holmes, 2019) to interact with the SRA database by building on top of its built-in modules for different E-Direct utilities. Specifically, we implemented a new *EFetchResult* and *EFetchAnalyzer* that work in tandem to request and parse metadata for the provided run IDs. The result is represented as a table where a single row corresponds to one SRA run and columns reflect all the metadata fields found in the obtained response. In order to keep track of different metadata levels (study, sample, experiment and run), we introduced a set of cascading Python data classes to delineate the hierarchical relationship of the SRA metadata entries (Fig. 3) and to preserve this structure in the final study metadata. Moreover, tight integration with QIIME 2's internal type validation system guarantees consistency of metadata generated by *q2-fondue* by ensuring the presence of all required metadata fields, as specified by NCBI.

2.2 Metadata retrieval

The *q2-fondue* package contains additional functions to simplify sequencing (meta)data retrieval and manipulation, particularly when multiple data fetches are necessary.

- ***get-all*** allows the simultaneous download of sequences and related metadata.
- ***merge-metadata*** concatenates metadata tables obtained from several independent *get-metadata* runs and allows the generation of a single, unified metadata artifact.
- ***combine-seqs*** merges sequences obtained from multiple artifacts obtained from several *get-sequences* runs (or from other external sources) into a single sequence artifact.
- ***scrape-collection*** retrieves accession IDs and associated DOI names from a Zotero web library collection (https://www.zotero.org) by using the *pyzotero* package (Hügel *et al.*, 2019). This enables seamless workflows for collecting IDs of interest from a literature collection, automatically downloading the data, and processing downstream with *q2-fondue* and QIIME 2.

3 Materials and methods

The *q2-fondue* plugin can be used to facilitate comparative analysis of any nucleotide sequence data and metadata available on the SRA. To demonstrate some example use cases, we used *q2-fondue* and QIIME 2 to analyze publicly available marker gene, whole genome sequence and shotgun metagenome data. All analyses described below are available as fully reproducible and executable Jupyter notebooks (available in the Data Availability section). These examples are provided merely as method demonstrations to showcase seamless integration/interoperation of *q2-fondue* with downstream analyses and do not represent complete analyses of biological data. Additional steps and larger comparative analyses would be required to derive meaningful conclusions, and to eliminate potential biases from covariates, which were not controlled for in this demonstration analysis.

3.1 Marker gene amplicon sequence data analysis

Marker gene amplicon sequencing (e.g. of 16S rRNA genes) is currently the most popular method for high-throughput, untargeted profiling of microbial communities as well as non-microbial diet metabarcoding and environmental DNA applications. To demonstrate the use of *q2-fondue* for comparative cross-study analysis of marker gene amplicon sequence data, we selected three studies that analyzed the development of the infant gut microbiome in distinct geographical locations: Lewis *et al.* (2017) with BioProjectID PRJEB16321, Davis *et al.* (2017) with BioProjectID PRJEB15633 and McClorry *et al.* (2018) with BioProjectID PRJEB23239. All three studies sequenced 16S rRNA genes in the V4 region with the forward 515F primer and a read length of 251 to 253 nucleotides. McClorry *et al.* (2018) additionally used the reverse primer R806.

The *get-all* action of *q2-fondue* was used to retrieve metadata and sequence data of all three studies from the SRA. To normalize metadata features of interest across studies, namely age and health status, we employed the Python library *pandas* (McKinney, 2010; Reback *et al.*, 2020). The downloaded single-read gene sequences were filtered according to the availability of metadata with the *q2-demux* QIIME 2 plugin (https://github.com/qiime2/q2-demux), denoised with the *q2-dada2* QIIME 2 plugin (Callahan *et al.*, 2016), https://github.com/qiime2/q2-dada2) and finally features were filtered by frequency, rarefied and summarized with the *q2-feature-table* QIIME 2 plugin (Bokulich *et al.*, 2018b) https://github.com/qiime2/q2-feature-table). Finally, the processed metadata and sequence data were used to train two Random Forest classifiers with *q2-sample-classifier* (Bokulich *et al.*, 2018a), https://github.com/qiime2/q2-sample-classifier) to predict an infant's age and its health status from its gut microbiome. The infant's age was reported in four binned age groups: 0–1, 1–4, 4–6 and older than 6 months. For the health status, an infant was identified as healthy if no disease-related features were reported, namely no indication of stunting, wasting, underweight, elevated C-reactive protein status, parasites or anemia. Both classifiers were trained and tested with 10-fold cross-validation using Random Forest classifiers grown with 500 trees. The performance of the trained classifiers was evaluated on the area under the curve (AUC) of the receiver operating characteristics (ROC) curve of the test set of each fold using *scikit-learn* implementations (Pedregosa *et al.*, 2011). The entire analysis can be reproduced by executing the *u1-amplicon.ipynb* Jupyter notebook, available in the Data Availability section.

3.2 Comparative whole-genome sequence data analysis

The initial list of SARS-CoV-2 whole genome sequencing data accession IDs was generated based on the metadata obtained from the Nextstrain.org platform (Hadfield *et al.*, 2018; https://data.nextstrain.org/files/ncov/open/metadata.tsv.gz, access date: February 08, 2022). Only records with available SRA accession IDs and the least amount of missing data (column *QC_missing_data* == '*good*') were retained. Furthermore, to limit the scope of this use case only those geographical regions were considered where enough samples were collected (at least 2000 samples per region). The *get-metadata* action of *q2-fondue* was used to retrieve metadata of 37 500 randomly sampled genomes (12 500 genomes per location) from the SRA. Obtained metadata was then supplemented with the original Nextstrain metadata by merging the two datasets on a common SRA run ID and only samples sequenced using single-end reads on the Illumina NextSeq 500/550 platforms were retained. Finally, sequencing data for 500 randomly sampled genomes (250 genomes per location) was fetched from the SRA using *q2-fondue*'s *get-sequences* method. Reads shorter than 35 nt were discarded using the *trim_single* method from the *q2-cutadapt* plugin (Martin, 2011; https://github.com/qiime2/q2-cutadapt) and the quality of the sequences was evaluated using the *summarize* action from the *q2-demux* QIIME 2 plugin (https://github.com/qiime2/q2-demux). MinHash signatures were computed for every sample and compared using the *q2-sourmash* plugin (*compute* and *compare* methods, respectively) (Ondov *et al.*, 2016; https://github.com/dib-lab/q2-sourmash). A t-SNE plot was generated from the resulting distance matrix using *q2-diversity* (*tsne* method with a learning rate of 125 and perplexity set to 18, https://github.com/qiime2/q2-diversity, Halko *et al.*, 2011) and visualized using *matplotlib* and *seaborn* Python packages (Hunter, 2007; Waskom, 2021). Finally, to determine whether MinHash signatures are predictive of SARS-CoV-2 geographic origin, k-nearest-neighbors classification with 10-fold cross-validation was applied through the *q2-sample-classifier* plugin (Bokulich *et al.*, 2018a; https://github.com/qiime2/q2-sample-classifier). The entire analysis can be reproduced by executing the *u2-genome.ipynb* Jupyter notebook, available in the Data Availability section.

3.3 Shotgun metagenome sequence data analysis

To fetch metadata for all the shotgun metagenome samples from the Tara Oceans Expedition (Tara Oceans Consortium Coordinators *et al.*, 2015) we used *q2-fondue*'s *get-metadata* action with the following BioProject IDs: PRJEB1787, PRJEB4352, PRJEB4419, PRJEB9691, PRJEB9740 and PRJEB9742. After the removal of missing values, the resulting metadata table was randomly sampled to 100 records and used as input to the *draw-interactive-map* action from the *q2-coordinates* plugin (Bokulich and Caporaso, 2018; https://github.com/bokulich-lab/q2-coordinates) to visualize values of the sensors used during the expedition according to their geographical location. Additionally, sequences corresponding to 10 samples at two different locations were fetched using the *get-sequences* action. To reduce the computational time required for this use case demonstration, the reads were subsampled to 20% of the original read count and the resulting artifact (containing single-end reads) was used to calculate and compare MinHash signatures (see the previous section). A PCoA analysis was performed on the resulting distance matrix using the *pcoa* action from the *q2-diversity* plugin (https://github.com/qiime2/q2-diversity; Halko *et al.*, 2011) and the PCoA plot was visualized using *matplotlib* and *seaborn* Python packages (Hunter, 2007; Waskom, 2021). The entire analysis can be reproduced by executing the *u3-metagenome.ipynb* Jupyter notebook, available in the Data Availability section.

4 Results

Any meta-analysis can be carried out using raw experimental data, its associated metadata or a combination of both. To demonstrate the versatility of *q2-fondue* in all those scenarios, and seamless integration/interoperability with downstream bioinformatics tools, we performed three example use case data analyses using amplicon, whole genome, and shotgun metagenome sequencing data and related metadata. All three use cases employ QIIME 2 plugins to process received data and illustrate how *q2-fondue* can immensely increase data analysis reproducibility and transparency by including details on the raw data fetching in the QIIME 2 provenance. These analyses are only meant to demonstrate some example use cases for *q2-fondue* and should not be interpreted as biologically meaningful results.

4.1 Use Case 1: amplicon sequence data analysis

As a demonstration of *q2-fondue*'s capacities in enabling the collection and analysis of amplicon sequencing data, we selected three infant gut microbiome development studies from distinct geographical locations: the study by Lewis *et al.* (2017) from Georgia, Davis *et al.* (2017)'s study from Gambia and McClorry *et al.* (2018)'s study from Peru. We used the BioProject IDs reported by those studies to fetch the corresponding raw metadata and sequencing data. This provided us with 350 sequence samples each annotated with 148 metadata features.

After performing filtering, normalization and denoising steps on the raw 16S rRNA gene sequences (see Fig. 4A for an overview of plugins and actions used throughout this use case), a total of 3880 amplicon sequencing variants (ASVs) were identified for 330 samples. The available metadata was used to define binned age groups. The distribution of samples per age group as well as the analyzed age range differ markedly between studies (Fig. 4B). We further defined a binary health status which denotes whether the sample stems from a healthy or unhealthy infant (see Methods for more details). Across all studies, 194 unhealthy and 136 healthy infant samples were identified. Figure 4C displays the fraction of healthy infants in each of the three geographic locations covered by the selected studies. It shows that the fraction of healthy infants is very different between studies and geographical locations with Lewis *et al.* (2017) having analyzed only healthy infants in Georgia and Davis *et al.* (2017) and McClorry *et al.* (2018) having analyzed mainly unhealthy infants in Gambia and Peru, respectively.

Finally, we trained two Random Forest classifiers with 10-fold cross-validation on the processed microbiome sequence data to predict the age group and health status of each sample, respectively. The classifiers were evaluated on the test set of each fold and revealed a better performance in predicting age groups (macro averaged AUC = 0.85, Fig. 4D) than health status (macro averaged AUC = 0.58, Fig. 4E). This initial result would, however, require further careful analysis as the individual studies differ in many variables (e.g. age range, health status and geographical location) which we did not account for in this demonstration. Hence, differences in predicted age bins and health status could be artificially inflated by the differences in geolocation or study design. The classifiers trained here might not be capturing age-specific or health-specific features but rather features stemming from the particular study setups.

4.2 Use Case 2: Whole genome sequence data analysis

To illustrate how q2-fondue can be used as an entry-point to analysis of whole genome sequencing data we turned to one of the most rapidly growing datasets of the recent years: the SARS-CoV-2 genome dataset. We used all of the pre-processed metadata obtained through the Nextstrain.org platform (Hadfield *et al.*, 2018) to identify samples that have been deposited in the SRA. We subsampled genomes of SARS-CoV-2 variants from two geographic locations: Europe and North America. We then fetched the corresponding SRA metadata using *q2-fondue*, which was used to prepare our final list of genomes. To simplify the analysis and reduce technical variability, we focused only on samples sequenced using single-end reads on the Illumina NextSeq 500/550 platforms. Following the quality control step, we used the sourmash tool to readily compare viral genomes to one another by computing their MinHash signatures (Ondov *et al.*, 2016). The resulting distance matrix was then used to generate a t-SNE plot visualizing how sampled genomes group together. Figure 5B shows that the samples taken at different locations group together to form distinct geographic clusters. Finally, we used k-nearest-neighbors clustering to quantitatively compare genome MinHash signatures to predict SARS-CoV-2 geographic origin (Fig. 5C). We found that it was possible to classify the SARS-CoV-2 origin with an accuracy of 92%. This result is intended as a simple demonstration of the capability of *q2-fondue* in conjunction with other QIIME 2 plugins but should not be considered as final. Additional factors may need to be taken into account to get a complete picture of the relationship between geographic location and viral genome MinHash signatures: other sequencing platforms should be included and more samples from other continents should be added, as well as a careful evaluation of covariates that could bias these results. An overview of plugins and actions applied in this use case can be found in Figure 5A.

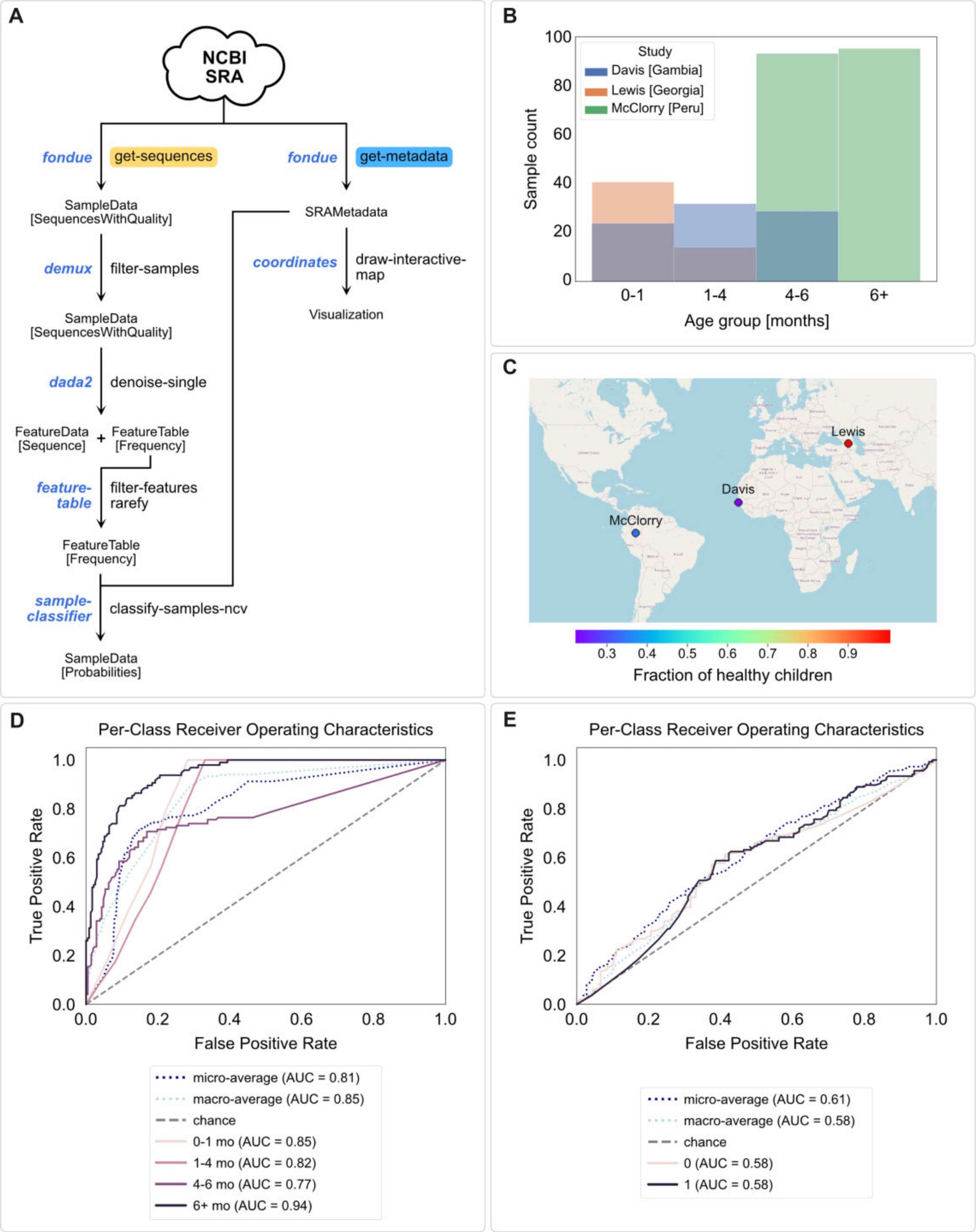


Fig. 4. Analysis of amplicon sequencing data from three infant gut microbiome development studies. (**A**) Overview of QIIME 2 actions used during the amplicon data analysis. (**B**) Counts of samples in the defined age groups per study. (**C**) Fraction of healthy infants in the geographic locations covered by the selected studies. (**D** and **E**) ROC curves of Random Forest classifiers predicting age groups (D) and health status (E), indicating better predictive accuracy for age groups (macro averaged AUC = 0.85) than health status (macro averaged AUC = 0.58) which is not to be trusted as the individual studies differ in many covariates that were not adjusted for in this demonstration

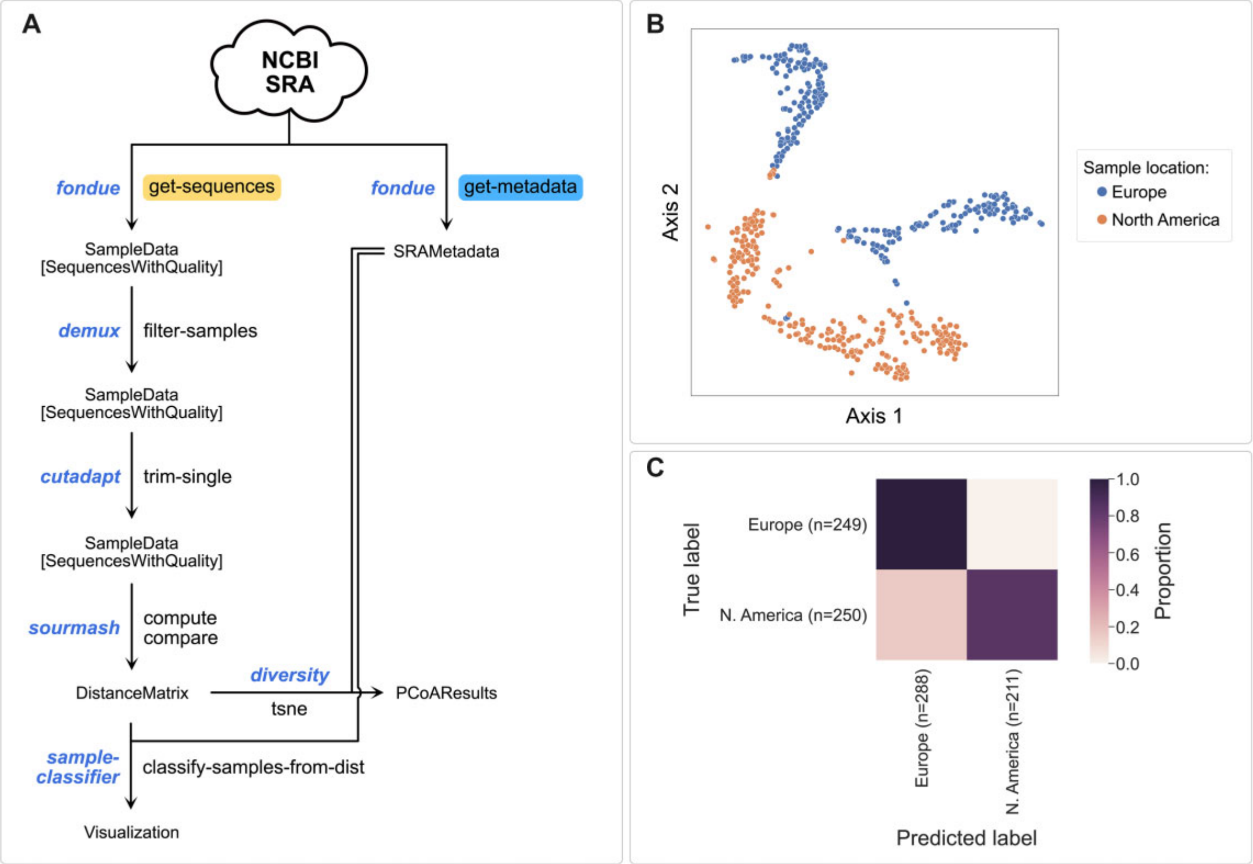


Fig. 5. Genome MinHash signatures are predictive of SARS-CoV-2 geographic origin. (**A**) Overview of QIIME 2 actions used during the genome data analysis. (**B**) t-SNE analysis of the SARS-CoV-2 genome MinHash distance matrix shows that virus samples can be grouped into distinct geographic clusters based on only genome hash signatures. (**C**) The same distance matrix can be used to reliably predict virus origin from genome hashes. K-nearest-neighbors clustering approach with 10-fold cross-validation was used to classify samples—a confusion matrix constructed from all test sets is shown

4.3 Use case 3: shotgun metagenome sequence data analysis

We used the Tara Ocean expedition dataset (Tara Oceans Consortium Coordinators *et al.*, 2015) to illustrate how geographic location included in sample metadata deposited in the SRA can be used to display sample properties, using *q2-fondue* and QIIME 2 (see Fig. 6A for an overview of plugins and actions used throughout this use case). We fetched metadata for six BioProjects containing 1049 ocean samples obtained through size fractionation followed by shotgun metagenome sequencing (Fig. 6B and C). As geographical coordinates of every sample are included in the SRA metadata, we could directly draw an array of interactive maps visualizing various sample properties using the *q2-coordinates* plugin (Bokulich and Caporaso, 2018). As an example, Figure 6D illustrates sample temperatures across the globe. Moreover, we randomly selected 10 samples collected at two distinct locations and used the corresponding sequences to calculate and compare their MinHash signatures. Using PCoA analysis of the resulting distance matrix, we could show that the samples can be separated by location when using only their genome hash signatures (Fig. 6E). More interactive visualizations can be found in the Jupyter notebook accompanying this manuscript (see Data Availability section).

4.4 Integration with QIIME 2 ecosystem

Since *q2-fondue* is a QIIME 2 plugin, it tightly integrates with and benefits from the rest of the QIIME 2 ecosystem. Sequences obtained through the *get-sequences* action can be directly passed into any other QIIME 2 action that accepts this data type (see Fig. 7 for an overview of actions applied in this study). In addition to defining format checks for SRA metadata objects, *q2-fondue* has implemented transformer functions to allow the metadata downloaded through the use of *get-metadata* action to serve as input to any QIIME 2 action that requires sample metadata. Furthermore, integration with QIIME 2's built-in provenance tracking system ensures that data fetching from the SRA is also included in the provenance graph (stored directly in all data outputs), enabling researchers to track and completely reproduce the entire analysis pipeline from data download to final visualizations.

5 Discussion

Declining costs and increasing throughput of nucleotide sequencing have fueled an exponential increase in published sequence data over the past two decades (Stephens *et al.*, 2015). These data have an immense reuse potential, which has led to a growing trend of sequencing-based meta-analyses (Fig. 1), paving the way to additional discoveries regarding general biological trends (Abbas *et al.*, 2019; Panagiotou *et al.*, 2013; Thompson *et al.*, 2017). However, such studies remain technically challenging and data acquisition and management are significant bottlenecks.

We developed *q2-fondue* to lower these hurdles, and to facilitate reproducible acquisition and management of metadata and nucleotide sequence datasets from the SRA (see Table 1 for a summary of the most important features). Its integration with the QIIME 2 framework offers complete provenance tracking of the entire process, multiple user interfaces, and thorough input/output data validation, allowing to conduct meta-analyses in a reproducible manner. Furthermore, *q2-fondue* outputs can be directly used with a wide range of QIIME 2 plugins, offering the user a smooth incorporation with any sequence-based analysis that is (or will become) available

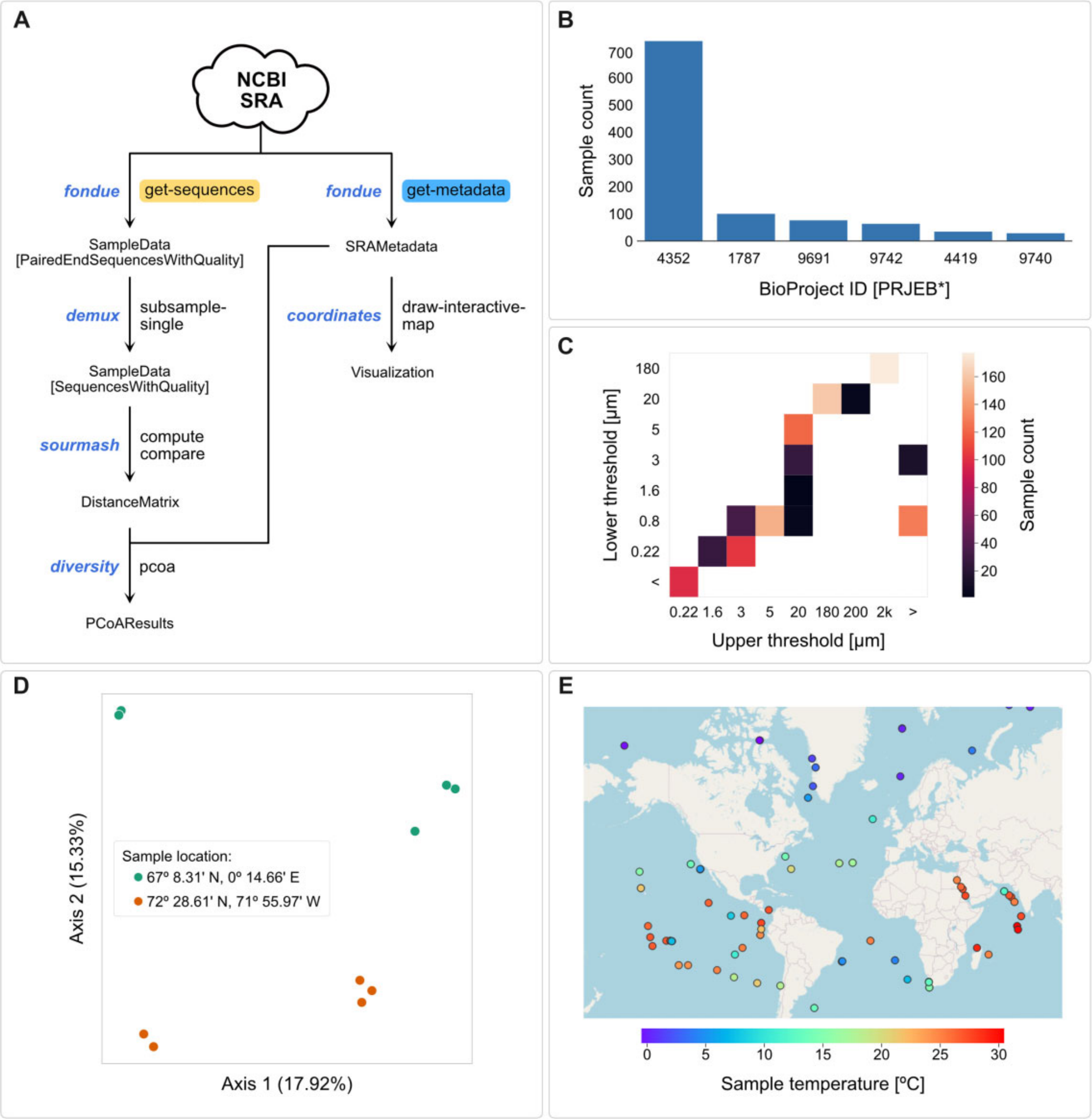


Fig. 6. Tara Oceans expedition (meta)data analysis. (**A**) Overview of QIIME 2 actions used during the metagenome analysis. (**B**) Counts of samples in the retrieved dataset according to BioProject ID. (**C**) Counts of samples corresponding to different fractions obtained through size fractionation. (**D**) PCoA analysis of the metagenome MinHash signatures of 10 samples taken at two randomly selected locations. Fraction of explained variance is shown for two plotted dimensions. (**E**) Temperature of samples taken at different geographical locations. Only 100 randomly selected samples are shown

within the QIIME 2 ecosystem. Even though the main target of the QIIME 2 framework is microbiome research, *q2-fondue* itself is intended to be agnostic to the research field and non-microbiome researchers can equally profit from its most important features (see Table 1) when performing downstream analyses without QIIME 2 (see tutorials at https://github.com/bokulich-lab/q2-fondue). Finally, *q2-fondue*'s integration with QIIME 2 offers users unparalleled support through the QIIME 2 forum—an exchange platform between users and plugin developers (with a current total of 5700 signed-up members).

Despite its ease of use, *q2-fondue* does not free the user of their due diligence in checking the details on the extracted datasets in the accompanying publications, where mismatches with obtained run metadata or sequences could be detected. The same holds for following best practices when performing meta-analyses or comparative analyses with data obtained using *q2-fondue*. The user must investigate sources of heterogeneity among individual studies included in the analysis (Thompson, 1994) and follow statistical procedures to ensure that the detected signals are not an artifact of the different study setups (Gurevitch *et al.*, 2018).

The *q2-fondue* demonstrations shown here represent only a few possible (although simplified) use cases for the software, and we envision many other possible applications for the analysis of diverse nucleotide sequence data types.

5.1 Future plans

The *q2-fondue* package remains under active development, and several additional functionality upgrades are planned in the future. As *q2-fondue* operates on large amounts of sequencing data we will

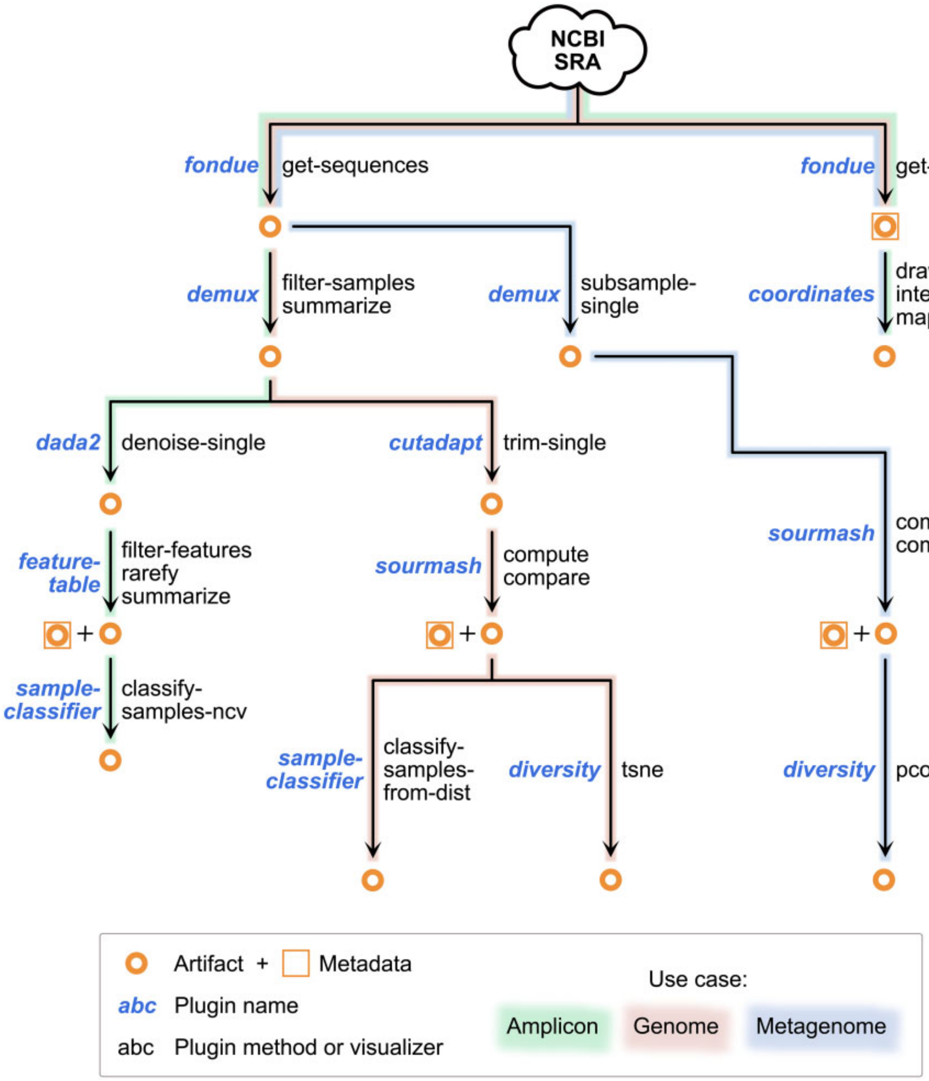


Fig. 7. Overview of *q2-fondue* integration with other QIIME 2 plugins and actions as applied in the three use cases presented in this study. This is only a limited demonstration of possible downstream uses for three different nucleotide sequence data types, not an exhaustive list

Table 1. Selection of the most significant issues faced by users when retrieving large amounts of sequencing data, together with their user-friendly solutions offered by *q2-fondue*

Problem	Solution offered by *q2-fondue*
Plethora of accession ID types complicates retrieval of sequences/metadata.	Conversion between different accession IDs is performed automatically. All associated parent and child accession IDs are recorded in the final metadata table.
Potential data loss on space exhaustion when fetching large amounts of runs.	*q2-fondue* keeps track of available disk space and will abort without data loss when the amount of space is insufficient.
Sequencing data requires pre-processing/name normalization before it can be used in downstream analyses.	*q2-fondue* takes care of renaming/standardizing all the files after retrieval.
Merged datasets and subsequent data analysis steps are not always reproducible.	Tight integration with QIIME 2 ensures that every data fetching and analysis detail is recorded in provenance stored together with every single output.
Diversity of metadata fetched from multiple studies complicates its application in subsequent analyses.	Metadata retrieved by *q2-fondue* is normalized into a single table with standardized columns.
Network issues and other errors lead to data loss and require cumbersome, repeated data fetches.	Data retrieval can be automatically repeated in case of encountered errors. In case of repeated failures, all errors are reported and can be investigated by the user once the download is finished. No data loss occurs.
Parallelization of custom SRA access scripts is complicated and time-consuming.	*q2-fondue* takes care of data retrieval/processing in a parallel way, making use of multiple threads and CPUs available on the user's system.

introduce several performance-enhancing updates that will allow better management of free storage space available during download as well as streamline downloading large numbers of accession IDs to avoid multiple re-fetches.

q2-fondue's metadata retrieval action already greatly simplifies downloading metadata of multiple projects and formatting those as a single result table. Several additional functions are planned to assist with the management and integration of diverse study and sample metadata. Moreover, we acknowledge that additional features may be needed, particularly for accessing non-microbial datasets (as the demonstrations in this publication focus on microbial datasets). We are motivated to further improve *q2-fondue* to encompass diverse use cases and invite feature requests via *q2-fondue*'s GitHub repository or the QIIME 2 forum.

Finally, to unlock the potential of sequencing data stored in and processed by other repositories we will add support for (meta)data retrieval from various other databases (e.g. MGnify; Mitchell *et al.*, 2020). Altogether, we hope that *q2-fondue* can become the tool of choice for interacting with the SRA and other similar repositories, while at the same time seamlessly integrating with the whole QIIME 2 ecosystem, hence enabling a wide range of available analysis types.

Acknowledgements

We thank Evan Bolyen (Northern Arizona University) for insightful discussions on metadata processing and working with the SRA and SRA Toolkit. We also thank Anton Lavrinienko (ETH Zürich) for his valuable comments on the manuscript.

Funding

This work was partially supported by the grant [#2021-362] of the Strategic Focus Area 'Personalized Health and Related Technologies (PHRT)' of the ETH Domain (Swiss Federal Institutes of Technology) and the Swiss National Science Foundation [310030_204275]. L.F. gratefully acknowledges the support of the Swiss Government Excellence Ph.D. Scholarships for foreign students.

Conflict of Interest: none declared.

Data availability

An introductory tutorial with background information and examples for the usage of *q2-fondue* can be accessed at https://github.com/bokulich-lab/q2-fondue/blob/main/tutorial/tutorial.md. All Jupyter notebooks used in this article are available under https://github.com/bokulich-lab/q2-fondue-examples.

References

Abbas,A.A. *et al.* (2019) Redondoviridae, a family of small, circular DNA viruses of the human oro-respiratory tract that are associated with periodontitis and critical illness. *Cell Host Microbe*, **25**, 719–729.e4.

Alföldi,J. and Lindblad-Toh,K. (2013) Comparative genomics as a tool to understand evolution and disease. *Genome Res.*, **23**, 1063–1068.

Amann,R.I. *et al.* (2019) Toward unrestricted use of public genomic data. *Science*, **363**, 350–352.

Baker,M. (2016) 1,500 Scientists lift the lid on reproducibility. *Nature*, **533**, 452–454.

Berman,F. *et al.* (2014) Building global infrastructure for data sharing and exchange through the research data alliance. *D-Lib Mag.*, **20**, https://doi.org/10.1045/january2014-berman.

Bokulich,N.A. *et al.* (2018) q2-sample-classifier: machine-learning tools for microbiome classification and regression. *J. Open Source Softw.*, **3**, 934.

Bokulich,N.A. *et al.* (2018) Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*, **6**, 90.

Bokulich,N. and Caporaso,G. (2018) *Nbokulich/q2-Coordinates: 2018.11*. Zenodo. https://doi.org/10.5281/zenodo.2124295.

Bolyen,E. *et al.* (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.*, **37**, 852–857.

Buchmann,J.P. and Holmes,E.C. (2019) Entrezpy: a python library to dynamically interact with the NCBI entrez databases. *Bioinformatics (Oxford, England)*, **35**, 4511–4514.

Callahan,B.J. *et al.* (2016) DADA2: high-resolution sample inference from illumina amplicon data. *Nat. Methods*, **13**, 581–583.

Choudhary,S. (2019) Pysradb: a python package to query next-generation sequencing metadata and data from NCBI sequence read archive. *F1000Research*, **8**, 532.

Davis,J.C.C. *et al.* (2017) Growth and morbidity of Gambian infants are influenced by maternal milk oligosaccharides and infant gut microbiota. *Sci. Rep.*, **7**, 40466.

Gurevitch,J. *et al.* (2018) Meta-analysis and the science of research synthesis. *Nature*, **555**, 175–182.

Hadfield,J. *et al.* (2018) Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, **34**, 4121–4123.

Halko,N. *et al.* (2011) An algorithm for the principal component analysis of large data sets. *ArXiv:1007.5510 [Cs, Stat]*. http://arxiv.org/abs/1007.5510.

Huang,Y. and Gottardo,R. (2013) Comparability and reproducibility of biomedical data. *Brief. Bioinformatics*, **14**, 391–401.

Hügel,S. *et al.* (2019) *Urschrei/Pyzotero: Zenodo Release*. Zenodo. https://doi.org/10.5281/zenodo.2917290.

Hunter,J.D. (2007) Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95.

Ioannidis,J.P.A. and Trikalinos,T.A. (2005) Early extreme contradictory estimates may appear in published research: the Proteus phenomenon in molecular genetics research and randomized trials. *J. Clin. Epidemiol.*, **58**, 543–549.

Kans,J. (2013) Entrez direct: e-utilities on the Unix command line. In: *Entrez Programming Utilities Help [Internet]*.National Center for Biotechnology Information (US), Bethesda (MD).

Katz,K. *et al.* (2022) The sequence read archive: a decade more of explosive growth. *Nucleic Acids Res.*, **50**, D387–D390.

Kim,Y.-M. *et al.* (2018) Experimenting with reproducibility: a case study of robustness in bioinformatics. *GigaScience*, **7**, 1–8. https://doi.org/10.1093/gigascience/giy077.

Kodama,Y. on behalf of the International Nucleotide Sequence Database Collaboration. *et al.* (2012) The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.

Leinonen,R. *et al.* (2011a) The European nucleotide archive. *Nucleic Acids Res.*, **39**, D28–D31.

Leinonen,R. *et al.*; International Nucleotide Sequence Database Collaboration. (2011b) The sequence read archive. *Nucleic Acids Res.*, **39**, D19–21.

Lewis,Z.T. *et al.* (2017) The fecal microbial community of breast-fed infants from Armenia and Georgia. *Sci. Rep.*, **7**, 40932.

Lloyd,K.G. *et al.* (2018) Phylogenetically novel uncultured microbial cells dominate earth microbiomes. *MSystems*, **3**, e00055–18.

Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.*, **17**, 10.

Mashima,J. *et al.* (2017) DNA data bank of Japan. *Nucleic Acids Res.*, **45**, D25–D31.

McClorry,S. *et al.* (2018) Anemia in infancy is associated with alterations in systemic metabolism and microbial structure and function in a sex-specific manner: an observational study. *Am. J. Clin. Nutr.*, **108**, 1238–1248.

McKinney,W. (2010) *Data Structures for Statistical Computing in Python*. pp. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a.

McNutt,M. *et al.* (2016) Liberating field science samples and data. *Science*, **351**, 1024–1026.

Meadows,J.R.S. and Lindblad-Toh,K. (2017) Dissecting evolution and disease using comparative vertebrate genomics. *Nat. Rev. Genet.*, **18**, 624–636.

Meyer,F. *et al.* (2008) The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.

Mitchell,A.L. *et al.* (2020) MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.*, **48**, D570–D578.

Ondov,B.D. *et al.* (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, **17**, 132.

Panagiotou,O.A. *et al.* (2013) The power of meta-analysis in genome wide association studies. *Annu. Rev. Genomics Hum. Genet.*, **14**, 441–465.

Parks,D.H. *et al.* (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.*, **2**, 1533–1542.

Pedregosa,F. *et al.* (2011) Scikit-learn: machine learning in python. In: *Machine Learning in Python*, Vol. **6**, 2826–2830. https://doi.org/10.48550/ARXIV.1201.0490.

Reback,J. *et al.* MomIsBestFriend (2020) *Pandas-Dev/Pandas: Pandas 1.0.3*. Zenodo. https://doi.org/10.5281/zenodo.3715232.

Reichman,O.J. *et al.* (2011) Challenges and opportunities of open data in ecology. Science , **331**, 703–705. https://doi.org/10.1126/science.1197962.

Serghiou,S. *et al.* (2016) Field-wide Meta-analyses of observational associations can map selective availability of risk factors and the impact of model specifications. *J. Clin. Epidemiol.*, **71**, 58–67.

Stephens,Z.D. *et al.* (2015) Big data: astronomical or genomical? *PLoS Biol.*, **13**, e1002195.

Pesant,S. *et al.*; Tara Oceans Consortium Coordinators. (2015) Open science resources for the discovery and analysis of Tara oceans data. *Sci. Data*, **2**, 150023.

(2019) The path to open data. *Nat. Rev. Nephrol.*, **15**, 521.

Thompson,L.R. *et al.*; Earth Microbiome Project Consortium. (2017) A communal catalogue reveals earth's multiscale microbial diversity. *Nature*, **551**, 457–463.

Thompson,S.G. (1994) Why sources of heterogeneity in meta-analysis should be investigated. *BMJ (Clinical Research Ed.)*, **309**, 1351–1355.

Waskom,M. (2021) Seaborn: statistical data visualization. *J. Open Source Softw.*, **6**, 3021.

Wilkinson,M.D. *et al.* (2016) The FAIR guiding principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.

Yilmaz,P. *et al.* (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.*, **29**, 415–420.

Youens-Clark,K. *et al.* (2019) iMicrobe: tools and data-driven discovery platform for the microbiome sciences. *GigaScience*, **8**. https://doi.org/10.1093/gigascience/giz083.

Zamkovaya,T. *et al.* (2021) A network approach to elucidate and prioritize microbial dark matter in microbial communities. *ISME J.*, **15**, 228–244.

Zhu,Y. *et al.* (2013) SRAdb: query and use public next-generation sequencing data from within R. *BMC Bioinformatics*, **14**, 19.